Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradoxes, and the 2016 US Presidential Election [1]

Summary by Thea Quistgaard*
(Dated: March 6, 2019)

The following is a summary of the article "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradoxes, and the 2016 US Presidential Election" by Xiao-Li Meng published in the Annals of Applied Statistics 2018 vol. 12[a]. The article contains many aspects of the Paradises and Paradoxes in Big Data, but I will only focus on the few points presented in the introduction.

The article presents a way to access both data quality and quantity and problem difficulty of a given (probabilistic or not) sampled data set. Furthermore it presents the seemingly statistical paradise of Big Data (as much data as you need, available any time you want it), as a pitfall to be wary of, or as Meng puts it "*The bigger the data, the surer we fool ourselves*".

An interesting question. A question which has put focus on the data quality-quantity trade-off is the following: "*Which one should I trust more: a 1% survey with 60 % response rate or a non-probabilistic dataset covering 80 % of the population?*". To be able to even try to answer this question we first need to find an identity which can link data quality, data quantity and problem difficulty. Meng presents the identity $\bar{G}_n - \bar{G}_N$ - the error on the sample mean of an estimator from the actual population mean. It is found that this error can be defined as

$$\bar{G}_n - \bar{G}_N = \frac{corr_J[R_J, G_J] \cdot \sigma_{R_J} \cdot \sigma_G}{f} \tag{1}$$

Here R_J is the so called recording mechanisms, where $R_j = 1$ for $j \in I_n$(I_n is a size n subset of $1, \ldots, N$, our population) and $R_j = 0$ otherwise. f is the relative sample size, $f = \frac{n}{N}$. σ_G is the standard deviation of the estimator and $\sigma_{R_J} = \sqrt{Var_J(R_J)} = \sqrt{f(1-f)}$.
This all results in the final identity:

$$\bar{G}_n - \bar{G}_N = \rho_{R,G} \cdot \sqrt{\frac{1-f}{f}} \cdot \sigma_G \tag{2}$$

Let us start from the last term: the term σ_G represents the problem difficulty, which is the variation over G in the sample. If G_J is a constant, then $\sigma_G^2 = 0$, whereas the more variation G has, the larger the problem difficulty.
The middle term, $\sqrt{\frac{1-f}{f}}$, represents the data quantity, and also reflects the relative size of the sample compared to the population, f. When the whole population is contained in the sample, this term renders zero error since f = 1. On the other hand it renders infinite error, when no data is recorded. And then we have the first term, $\rho_{R,G}$. This is the most critical part of the product, as it captures data quality. This term is the *data defect correlation* and is the correlation, $corr_J[R_J, G_J]$, between the variablr X_j and the response/recording indicator R_j. This indicator is present to indicate how the data was sampled. $R_j = 1$ for $j \in$ sample and 0 otherwise. The data defect correlation captures data quality since it measures both the sign and degree of selection bias caised by the R-mechanism(method of sampling e.g. probabilistic or non-probabilistic). If larger values of G have a tendency to be either more or less recorded then the value $\bar{G}_n$ either over- or underestimates the tru value $\bar{G}_N$. The degree of bias is captured by the size of $\rho_{R,G}$ and the over-/underestimate is captured by the sign.
Statistically, we can use this model, if the recorded values of G are trustworthy. This means that if a respondent answers with a certain response (ex.: "Vote for Clinton"), then it means only that the respondent is sufficiently inclined to vote for Clinton at the time of the response and nothing else. Otherwise we will be dealing with response bias which would complicate our problem even further!

MSE of $\bar{G}_n$. Now we can express the root mean-squared error (MSE) of $\bar{G}_n$(under any R-mechanism) to find out how to reduce it.

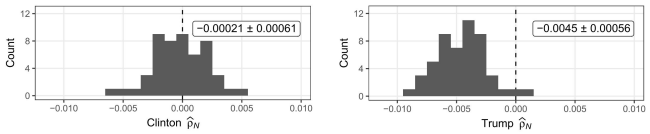
$$\begin{aligned} MSE_R(\bar{G}_n) &= E_R[\bar{G}_n - \bar{G}_N]^2 \\ &= E_R[\rho_{R,G}^2] \cdot \frac{1-f}{f} \cdot \sigma_G^2 \\ &\equiv D_I \cdot D_O \cdot D_U \end{aligned} \tag{3}$$

This subsequently gives us three ways of reducing the MSE:

- **Increase data quality** by reducing $D_I = E_R[\rho_{R,G}^2]$ the *Data Defect Index (d.d.i.)*.
- **Increase the data quantity** by reducing the Dropout Odds, $D_O = \frac{1-f}{f}$.
- **Reduce the difficulty** of the problem by reducing the Degree of Uncertainty, $D_U = \sigma_G^2$. Typically only possible by adding additional information to the problem.

The article further shows that the most effective way to reduce the MSE is by increasing the data quality - which might be contrary to our standard belief of *the more the*

* ksr966, Thea.Quistgaard@gmail.com
[a] https://statistics.fas.harvard.edu/files/statistics-2/files/statistical_paradises_and_paradoxes.pdf



(a) Clinton poll[3]. (b) Trump poll[3].

data, the better the statistics. Thus it seems obvious to look at the data defect index (d.d.i.).

Data Defect Index. Some questions raised about the d.d.i. that need to be answered are the following: (1) "*What are the likely magnitudes of D_I when we have probabilistic samples?*" and (2) "*How do we calculate or estimate D_I for non-probabilistic data?*".
To answer (1) first: For any Simple Random Sampling (SRS) we have that $\bar{G}_n$ is unbiased for $\bar{G}_N$ and its MSE is the same as its variance,

$$V_{SRS}(\bar{G}_n) = \frac{1-f}{n} S_G^2, S_G^2 = \frac{N}{N-1}\sigma_G^2, \tag{4}$$

which leads to the d.d.i. for any SRS to be given by

$$D_I \equiv E_{SRS}[\rho_{R,G}^2] = \frac{1}{N-1}, \tag{5}$$

It shows that for any probabilistic sampling $D_I \propto N^{-1}$ holds in general and will thus deisappear for large N - as we are used to.
Answering (2) is a little less mathematical and more applied statistical. The d.d.i. is something that needs to be evaluated individually for each problem. One issue with the d.d.i. is that it is not possible to estimate D_I from the sample itself without any assumption or knowledge about the R-mechanisms(recordings). Alas this is often the case, and thus D_I is unknown. But what can be done, proposed by Meng, is to ascertain an actual error and $\rho_{R,G}$ after an event and use this information on further likewise events. This will make it possible to construct a reasonable prior for $\rho_{R,G}$ or D_I from historical or neighbouring studies. For the 2020 US presidential election it will thus be possible to construct a plausible prior distribution of $\rho_{R,G}$ from histograms of prior $\rho_{R,G}$, see Figure **??** for illustration of data defect correlation from 2016 US presidential election.

A Law of Large Populations. When sampling probabilistically a central driving force for stochastic behaviors of the sample mean(and many other variables) is the sample size n. We recognize this from the Law of Large Numbers and from the Central Limit Theorem. Meng shows that when the sampling is no longer probabilistic n is no longer the driving force, but the population size N is. This can be seen by looking at the z-score, which is still just the error but expressed in terms of a (probabilistic) sample variance.

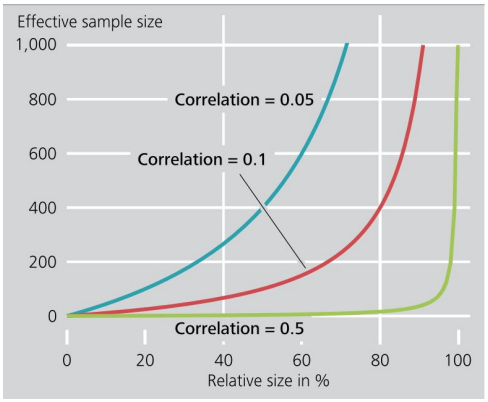


FIG. 2: Illustration of n_{eff} compared to the relative size.[a]

[a] Figure from Mehrhoof (2016)[2]

$$\begin{aligned} Z_{n,N} &\equiv \frac{\bar{G}_n - \bar{G}_N}{\sqrt{V_{SRS}}} \\ &= \frac{\rho_{R,G}\sqrt{\frac{1-f}{f}}\sigma_G}{\sqrt{\frac{1-f}{n}\frac{N}{N-1}\sigma_G^2}} \\ &= \sqrt{N-1}\rho_{R,G} \end{aligned} \tag{6}$$

This is of course not entirely correct as the actual MSE can be very different from $V_{SRS}(\bar{G}_n)$ but it is good enough for an estimator. This all leads to **Law of Large Populations (LLP)** which states that *Among studies sharing the same (fixed) average data defect correlation $E_R[\rho_{R,G}] \neq 0$, the stochastic error of $\bar{G}_n$, relative to its benchmark under SRS, grows with population size N at the rate of $\sqrt{N}$.* This shows us that we are under a curse of large populations as the error will grow with the population size.

The Return of the Monster N. To show how much damage a seemingly small data defect correlation can inflict, Meng has computed the effective sample size n_{eff} of a Big Data set by equating $MSE(\bar{G}_n)$ to the mean-squared error of the SRS estimator with the sample size n_{eff}. This leads to an expression for the effective sample size as

$$n_{eff} \leq \frac{f}{1-f} \cdot \frac{1}{D_I}. \tag{7}$$

An illustration of the effective sample size compared to the relative size, $f = \frac{n}{N}$, can be seen in Figure 2.

Real world application. This theory can help understanding and predicting the outcome of non-probabilistic sampling, such as elections. In the article, Meng has used data from the 2016 US Presidential Election to demonstrate the effectiveness of this theory. It helps guide us to an understanding of why some polls predict a much different outcome than what is actually seen under elections.
All in all this article presents the seeming statistical paradise of Big Data as a paradox more than a paradise and to tread carefully when venturing into the great unknowns of Big Data and non-probabilistic sampling.

[1] MENG, X.-L. (2018), Harvard University *Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election*, The Annals of Applied Statistics, 2018, Vol. 12, No 2, 685-726.

[2] MEHRHOFF, J. (2016). Executive summary: Meng, X.-L. (2014), "A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it)". Conference handout.

[3] MCDONALD, M. P. (2017). 2016 November general election turnout rates.